ARMY RESEARCH LABORATORY

Mixed Methodology to Predict Social Meaning for Decision Support

by Barbra E. Chin, Candace C. Ross, and Michelle T. Vanni

ARL-MR-0850 **September 2013**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0850 **September 2013**

Mixed Methodology to Predict Social Meaning for Decision Support

Barbra E. Chin, Candace C. Ross, and Michelle T. Vanni
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2013	Final	13 May 2013 to 5 July 2013

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Mixed Methodology to Predict Social Meaning for Decision Support	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Barbra E. Chin, Candace C. Ross, and Michelle T. Vanni	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Research Laboratory ATTN: RDRL-CII-T 28000 Powder Mill Road Adelphi, MD 20783-1197	ARL-MR-0850

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Army Research Office PO Box 12211 Research Triangle Park NC 27709-2211	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

For analysts of social content in language, popular Internet forums provide ample material for observing usage variation patterns. A single phenomenon, scrutinized through the lens of an established theoretical construct, focuses an approach to analysis, which is computationally tractable and exploitable. The contribution to ongoing work reported on here shows how diverse and complex manifestations of a style-switching variant of the code-switching phenomenon can serve profitably as data input to machine learning. On a group membership prediction task, logistic regression results for user posts containing style features were in the high 80s. A novel representation of structures in posts without style features boosted results for this group as well. We indicate how this approach may extend to popular social media sites, such as Facebook, to inter-language code-switching and diverse computational tasks.

15. SUBJECT TERMS

Code-switching, African language, machine learning, social media

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michelle T. Vanni
Unclassified	Unclassified	Unclassified	UU	30	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-0367

Contents

List of Figures

List of Tables

1. Introduction

Analysts of social meaning in language rely on the post data of Internet forums to provide the material for observing linguistic variations that pattern within and across non-linguistic user, post, or topic categories (threads). An analytical approach that employs a single socio-linguistic phenomenon, such as *code-switching* or its variant *style*-switching, can function as a window into linguistically conveyed social content, such as user identity, group membership, or political affiliation, associated with a given category of user, topic, or post.[1]

In computational modeling of sociolinguistic content in digital text, scientists depend on general features, such as unigrams, n-grams, punctuation, and parts of speech, found in data associated with a given content category. Features with a high degree correlation receive proportionally more weight. An approach to modeling that considers as well-known content-specific features, such as the style- or code-switching (CS) associated with a given group, political position, or sentiment, can enhance the accuracy of applications employing the model.

Here we report on a contribution to ongoing work that highlights the promise of pairing text and style features with training algorithms for predicting social meaning, such as user identity and affiliation, which are relevant for Soldier and Army decision-making purposes. Our aim is to show the promise of a novel two-prong approach to identifying social content in language. Starting with section 4 on methods, the report organization is twofold. First, we present a qualitative analysis of online forum data, based on a model proposed by Myers-Scotton (1983). Next, we show quantitative analysis and predictive modeling and how it is enhanced with style features in LightSIDE (Mayfield and Rosé, in press). The qualitative focus is on language patterns associated with identity; the quantitative focus is on the presence or frequency of text and style features in patterns. Both work toward a central goal: to exploit linguistic cues in determining social identity and inform effective trust-dependent Army decisions.

2. Background

The analyses and findings in this report had their origins in Dr. Carolyn Rosé's interdisciplinary classification research. At the Carnegie Mellon University (CMU) Language Technologies Institute (LTI), Rosé directs large-scale text classification studies, one of which uses as data postings to a Web forum for members of street gangs. The online forum is known as HoodUp.[2]

[1] To Haugen (1953) goes credit for coining the term *code-switching* as "introduction of a single unassimilated word from one language into another." Scholarly definitions proposed since share an idea, which we adopt here: effortless conversational alternation between languages, dialects, or language varieties. We thus view *style*-switching as a form of code-switching.

[2] The online gang forum's Web address is http://www.thehoodup.com.

The project investigates (1) the use of gang language, identified through members' characteristic style features; and (2) the correlation between language use and gang affiliation or social factors relating to gangs.[3]

Throughout HoodUp, the language in postings ranges from a strong use of the vernacular characteristic of African American English (AAE) to regular usage of Standard American English (SAE) that also ranges in use of stylistic features that identify users as members of certain street gangs, inter-gang *alliances*, or intra-gang *sets*. This style usage is heavy enough to be considered a form of CS. Examining this language, analysts seek to understand and characterize how identity functions in this type of code switching. Put another way, they look for clues to motivations behind style usage, and to the extent possible, they systematize the evidence supporting their analyses.

Systematization allows for automatic extraction of style features. To correlate the language use and gang affiliation, we use LightSIDE, a technology that extracts general text features, e.g., unigrams, binary n-grams, and punctuation, from the posts of specific users. The values or counts for each forum user's posts are made available to the machine learning algorithms in LightSIDE. Using features and algorithmic options, we can build a statistical model that will predict a user's gang membership based solely on their use of language.

While aspects of gang language, such as the stylistic tendencies of the language of graffiti (Adams and Winter, 1997) and the linguistics of gang speech (Conquergood, 1994; Garot, 2007), have been the subject of wide-ranging research on gangs, the stylistics of gang language online, as a mode of code switching that reflects the infrastructure of the larger gang community, has been little studied. Even the stylistic norm detailed by Nguyen and Rosé (2011) to explore the online expression of self would be able to accommodate the stylistics of the language used in the multifaceted lifestyle of gangsters only with considerable expansion.

Gang infrastructure is variable, regional, and indirectly related to likelihoods of violence. In densely populated areas, intra-gang groups known as sets figure in the structure; denseness also means less territory for gang occupation; the "turf" value increases, multiple gangs claim a block and the likelihood of violence rises. In sparsely populated areas, infrastructure carries through a "wide turf," or multiple blocks; a gang entering another gang's clearly defined territory is an undeniable violation and, again, there's an increased probability of violence. Note also that the violence can often be intra-gang. Superior gang members inflict violence upon their lesser members, who make attempts at other gangs' spaces, to maintain inter-gang peace.

Beyond infrastructure, and directly related to our study, are gangs' stylistic tendencies. Gang language, notably expressed as graffiti on city walls, includes names, threats, and expressions of disrespect, solidarity, and affection. Names are probably most recognized since symbols and

[3] The effort described here supports the project, Extracting Social Content in African Language and English (E-SCALES). The U.S. Army Research Office (ARO) Partnership in Research Transition (PIRT) program at ARO sustains this 3–5 year center initiative between Howard University (HU) and CMU.

names on walls delimit a gang's territory. On another gang's turf, they are an insult, a claim to superiority and ability to infiltrate. Yet, graffiti also expresses fondness, as with men boasting of female counterparts or honoring fallen fellow gangsters. Multiple names on an amatory wall show solidarity, and person names together show interpersonal connections within a larger gang structure. Virtual gang language parallels physical gang graffiti. In its complexity abides the cognitive and relational engine powering this endogenous nation within our society.

In negotiating and committing resources, Army leaders run great risks. This is especially true when processes involve decisions to trust individuals in native populations. As with the gang nation in the U.S., an adversarial organization in a strategically important country will incorporate characteristic style- or code-switching in the language they use. Research in detection of style-switching, CS, and associated motivations may well serve to mitigate the Army's trust-dependent risks with eventual development of systems for predicting a native individual's identity, affiliation, or intention, based on language alone.

3. Data

While all HoodUp user postings contribute to our research, the gang language project—which has already explored the influence of education and reasons for discontinuation—has selected for study eight of the hundreds of gangs represented. The language of these Web postings is the lens through which we observe gang members' relationships with each other and with the larger society. We are grateful for Adams and Winter (1997), who documented specific characteristics of gang language style and linked them to the language of graffiti and other written gang language.

HoodUp has over 1M posts and over 12K active users (Jeblee, Piergallini, Rodriguez, and Vaughn, 2013). Dr. Rosé's group's robust *Chive* Structured Query Language (SQL) database indexes forum users by both `<userName>` and unique `<userId>`, postings by `<postId>` and topic threads by `<threadId>`. Table 1 shows a sample of the data identification entries.

Table 1. Sample of data identification.

Username	User ID	Thread ID	Post ID	Post
tupacback	35541	54130	898150	This the new generation of gangstas haha. It's crazy seeing dudes you came up w/ and grew up w/ doing it like this haha.
MaCKKTR3Y83HCG	14870	15609	220311	watz gr38vin crim. W3r uk frxm hxmie?

This organization makes individual posts easily identifiable for qualitative analysis. We accessed posts, referred back to previously analyzed posts, and made comparisons with new posts via index queries to the Chive interface, pictured here as figure 1.

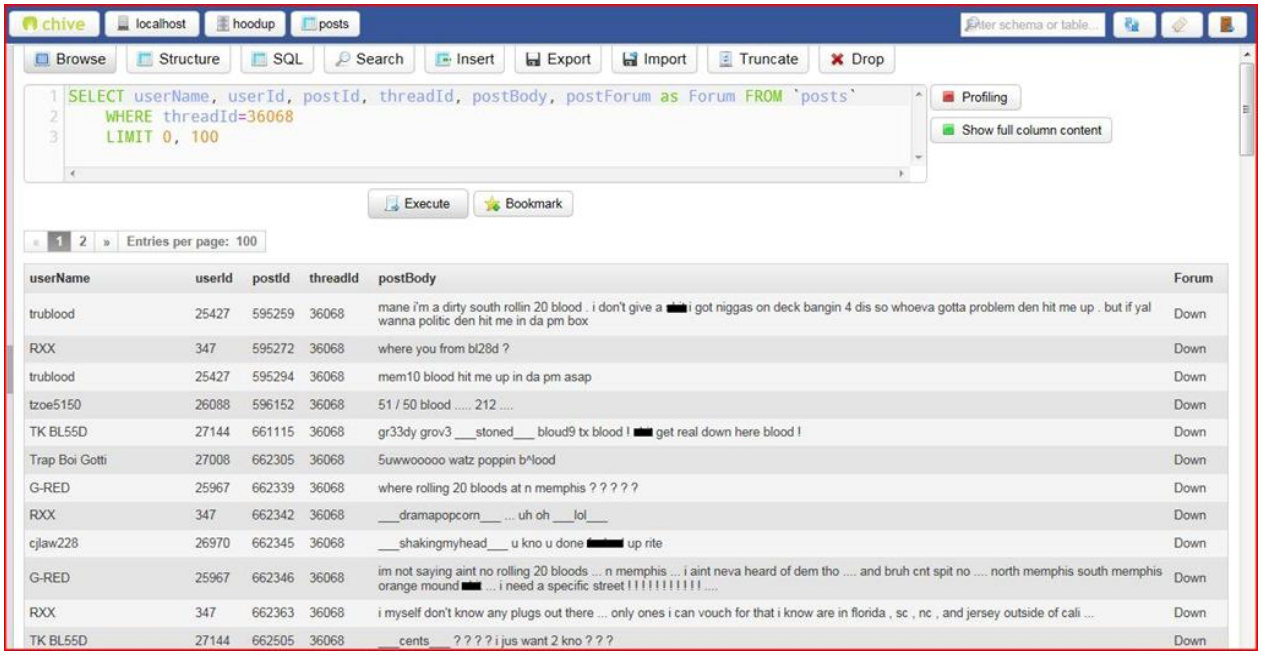


Figure 1. Interface to database for accessing posts.

The data used for quantitative analysis and computational modeling consisted of the extracted postings of forum users who had in some way defined their affiliation as being one of the eight specific gangs.[4]

The primary data element in Chive is the user's actual post. Transferred from the forum, the post is stored with the Chive-assigned and indexed identification one-ups mentioned previously, along with associated Web-site-dependent metadata elements, such as `<subforum>` and `<time>`, which are also transferred and stored along with the post.

Chive's HoodUp data are stored in multiple files. Note that Dr. Rosé's group had previously named, defined, and recorded the occurrence of about 30 specific gang style features. Table 2, taken from Jeblee et al. (2013), shows examples of these style features. They include placement of a caret next to a gang name's initial letter as well as substitutions of various types. They had also, for each user, concatenated the body of all posts, calculated the number of occurrences of each feature, and stored the counts in a comma separated value (CSV) file in which the column headers reflected the style feature.[5]

[4] When users avoid self-identifying as members of a specific gang, we approximate definitive ground truth as nearly as possible with double-blind human annotation and a report of the K agreement coefficient.

[5] A CSV file permits a simple matrix display of the data without special features, as in an Excel workbook.

For each user row, there are 28 pre-calculated features, one per column, across the eight gangs. Examples of these column features include <3eFreq>, which is a count of the "e → 3" feature from table 2, a substitution of the number '3' in place of the letter 'e' in words and <CrabFreq>, which counts the occurrence of the word 'crab', an insult used against Crips. Thus, files are of varying size, depending on style usage and user counts. While one major file contains posts and metadata of almost 1800 members, who were heavy users of gang style features, a much smaller file contains posts and metadata of the only about 800 users whose incorporation of style features is negligible.

Table 2. Sample of gang style features observed in HoodUp (from Jeblee et al., 2013).

Style Feature	Origin or meaning
b^, c^, h^, p^	"Bloods up" Positive towards Bloods, Crips, Hoovers, Pirus, respectively
b → bk, c → ck h → hk, p → pk	Blood killer, Crip killer Hoover killer, Piru killer
b → c	Replace 'b' for Blood with 'c' for Crip
c → b	Replace 'c' for Crip with 'b' for Blood
o → x, o → ø	Represents crosshairs, crossing out the '0's in a gang like Rollin' **60s** Crips
b → 6	Represents the six-pointed star. Symbol of Folk Nation and the affiliated Crips.
e → 3	Various. One is the trinity in Trinitario.
s → 5	Represented the five-pointed star. Symbol of People Nation and the affiliated Bloods.

A central problem and a focal point of our quantitative modeling efforts was the automatic classification of users who avoided incorporating such gang style features in their posts. Data developed to respond to this research question are described in section 4.2.3.

4. Methods

Our contribution's two-prong approach to determining social content in language requires (1) focus on a single linguistic phenomenon, such as CS; (2) focus on a single social meaning,

such as identity and group affiliation; and (3) concomitant qualitative analysis and quantitative modeling, with flexible feature development and testing. The method transfers readily to new types of CS, diverse social and interpersonal relations, as well as contrasting types of interaction. These can diverge on the physical versus virtual plane or the group versus one-on-one plane. In this way, the method shows great promise for development of generalized decision-support applications.

4.1 Qualitative

After developing background knowledge of the structure and history of street gangs such as the Crips, Bloods, Latin Kings, Hoovers, and Trinitarios, among others, we qualitatively examined the use of language across various threads and posts. Following the posts of a single forum member, within the framework described in this section, often provided valuable intuition about the member's specific use of style features. That led to insights into how identity functioned within this type of code switching.

Myers-Scotton's (1993) *markedness model* provided the framework in which we qualitatively analyzed the code switching on HoodUp. According to this model, conversational participants have shared expectations about code choices and their communicative intentions, which are defined as expected or *unmarked*. This "communicative competence" entails an awareness of contextual acceptability and the extent to which linguistic choices are to be interpreted in a given context. *Marked language*, on the other hand, is unexpected and conveys more specific information about the use and intent of language, especially as a function of identity.

In this model, marked and unmarked choices fall along a continuum of being more or less marked, with unmarked choices being dominant, as they are the conventionalized and expected choices. This choice changes based on context and can even include code switching. As situation and context are critical to understanding whether language is marked or unmarked, we relied on thread topics and language use within individual posts as well as in complete threads to provide insight into the user's expression of identity. Using the markedness model, language as an identity function is central to how each thread and post is interpreted in the qualitative analysis.

4.2 Quantitative

For quantitative analysis, we employed the LightSIDE software's extraction and machine learning capabilities. Extracting data features and creating new data types, we trained predictive models with its logistic regression and support vector machine (SVM) engines. We outline these capabilities here.

4.2.1 Selecting Quantitative Characteristics: Extracting Features

LightSIDE feature extraction begins with upload of a corpus file in CSV format. Users designate the *text* field, the column from which features are extracted; and the *class* field, the column containing the values to be predicted. *Column features*, described in section 3, may also be

designated. General features, such as *unigrams*, line length, and regular expressions, can also be extracted.[6] These features can be used in conjunction with the column features to build numerous models. Figure 2 shows a screenshot of LightSIDE's *Extract Features* tab.

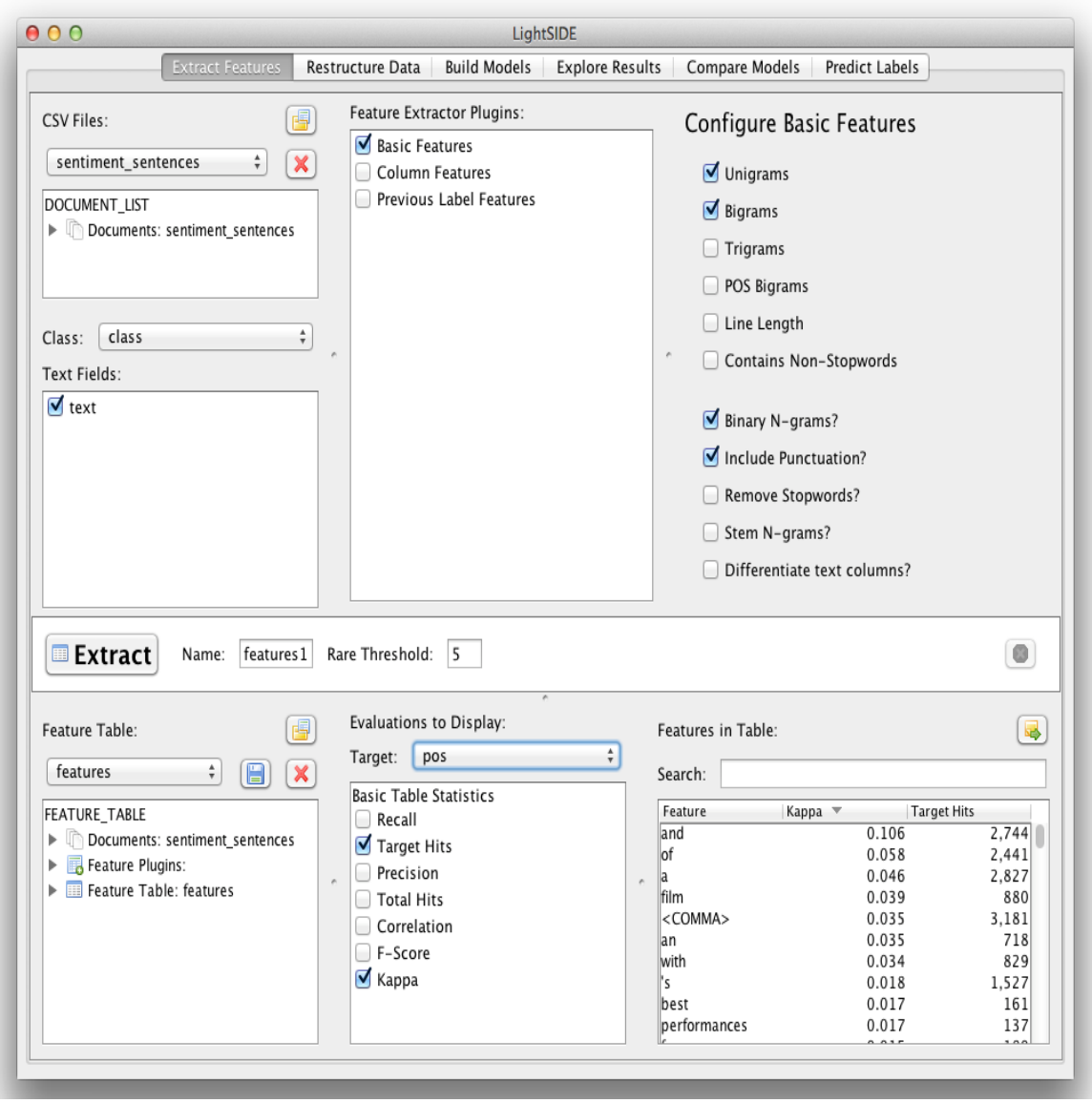


Figure 2. Screenshot of the LightSIDE software's *Extract Features* tab.

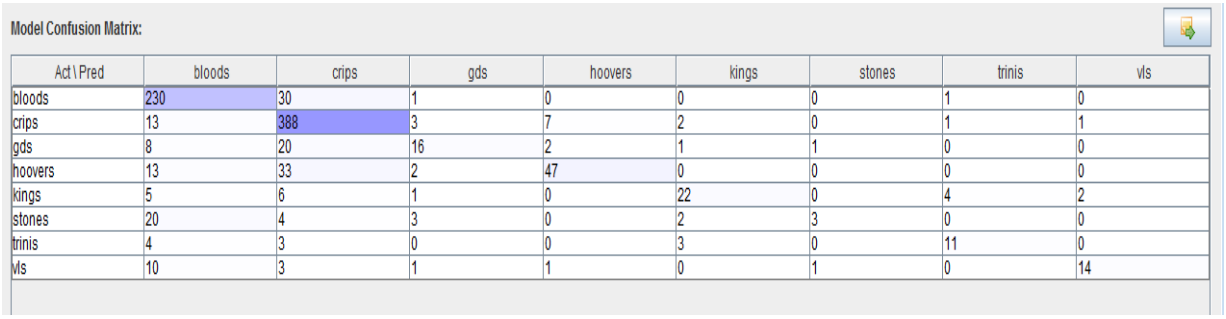
4.2.2 Applying Machine Learning to Features: Building Models

In LightSIDE, model building is facilitated by multiple algorithms and options within each for performance tuning. To model the HoodUp data, we used logistic regression with and without style features. We produced numerous models with varying accuracies by feeding the extracted

[6] Unigram feature types are simply character strings separated by white space and extracted from the text, e.g., names, acronyms, abbreviations, titles, numbers, and times. LightSIDE extracts every word and gives it a weight toward each gang, based on its relative usage in the posts of that gang's members. The same is true of regular expressions.

data to the algorithm. While on models built with style features, group membership prediction accuracy rose considerably, our goal is to create predictive models that are as accurate as possible.

To improve model accuracy, we try and determine which features are particularly contributive to the model's strengths and shortcomings. Observing feature weights and frequencies of occurrence, we noticed that infrequent users of style features were often misclassified. In fact, users on whose posts prediction accuracy was lowest were those whose posts were without any gang-specific style features. Subsequent efforts thus shifted to training and analyzing models on the corpus of posts authored by those ~800 gang forum users. We used LightSIDE's confusion matrix display capability to examine how the model operated and to determine where to focus to improve its accuracy (figure 3).



Model Confusion Matrix:

Act \ Pred	bloods	crips	gds	hoovers	kings	stones	trinis	vls
bloods	230	30	1	0	0	0	1	0
crips	13	388	3	7	2	0	1	1
gds	8	20	16	2	1	1	0	0
hoovers	13	33	2	47	0	0	0	0
kings	5	6	1	0	22	0	4	2
stones	20	4	3	0	2	3	0	0
trinis	4	3	0	0	3	0	11	0
vls	10	3	1	1	0	1	0	14

Figure 3. Confusion matrix for prediction of forum user affiliation with one of eight gangs.

The predictions displayed in figure 3 are for the approximately 800 users without style features in their posts. Note that the vertical axis represents the ground truth, the user's `'Actual'` gang; the horizontal axis is what the model `'Predicted'` as the user's gang affiliation, based on the posts. For example, the value `'13'` in the second row means that 13 actual Crips were predicted to be or classed as Bloods. This concept carries forth throughout the matrix and raises questions of how to tease apart the error-causing features in the misclassified data and how to reframe them for predictive accuracy.

4.2.3 Stretchy Patterns and Their Use in Models

The results of the classification of non-style-feature users in figure 3 show that logistic regression using generalized text features was able to correctly classify 0.75 of represented users. In observing the posts of the remaining 0.25, we performed error analysis by observing how linguistic elements in the posts compared and contrasted with those of users to whose gang they had been assigned. The first contrast observed was use of the name of one's own gang, e.g., `'bloods'` or `'trinis'`. In the feature distributions for previously built models, frequently occurring and heavily weighted feature words in the posts of members of a given gang had been grouped together into a category of terms and phrases associated with that gang. A few of these terms are displayed in table 3.

Table 3. Terms associated with each of eight gangs, as observed in member postings.

Bloods	**Crips**	**Gangster Disciples**	**Hoovers**	**Latin Kings**	**Black P. Stones**	**Trinitarios**	**Vice Lords**
blood	Crip	GDs	Hoova	Latin King	BPS	Trini	Vice lord
Piru	Locc	Gangsta Disciple	Hoover	Latin Queen	Stone	3ni	vls
Blxxd	Loccette	Black Gangsta Disciple	Hxxva	Almighty Latin King Nation	5tone	Trinitarios	Almighty Vice Lords
United Blood Nation	Gangster Crips	Maniac Gangsta Disciple	Hoover Gangsters	Latin King/ Queen Nation	People Nation	3nitario	Conservative Vice Lords

For the majority of users, these terms alone, properly weighted, along with accompanying features, correctly predicted affiliation. We inspected the posts of the incorrectly classified users for identifying characteristics beyond single words or phrases; we wanted to represent structurally any patterns found. That way, the patterns could also serve as features. Recognizing that term sets constituted gang name type categories, we observed as well recurring lexeme types—lemmas expandable into tokens—which combined in various ways with categorized and uncategorized text. These characteristics allowed us to call upon a feature construction capability of LightSIDE that Gianfortoni, Adamson, and Rosé (2011) found effective for binary classification.[7]

The capability, dubbed *stretchy patterns* (SPs), is appropriate for feature engineering when a specific word type category order recurs in a corpus. The ordering pattern is called "stretchy" because it accounts for the interspersion of a limited amount of uncategorized text. This material, which in practice constitutes its own category, is referred to with the term *Gap*. The SP capability has been defined as "a sequence of categories, which must not begin or end with a Gap category" (Gianfortoni et al., 2011: 53).

SPs can be thought of as word-level regular expressions. Where regular expressions are used to search text for patterned character sequences, we use SPs to search text for patterned word sequences. Words are assigned to specific LightSIDE word type categories, either pre-set or user-generated. For instance, the two phrases, *"Carl jogged around the track"* and *"Mary jogged slowly on the banked track"* can each be defined as the SP "`[PERSON-NAME]` jogged `[GAP]` track."[8] Table 4 shows examples of standard sentences expressed as SPs. SPs are particularly relevant for analysis of gang language, which makes creative use of words and expressions with standard meanings. Sequence matches account for context while the Gap mechanism provides flexibility and robustness well suited to linguistic variation.

[7] A Gianfortoni et al. (2011) evaluation of the stretchy pattern feature construction capability demonstrated significant improvement over standard baselines when used in modeling variation related to gender in personal weblogs.

[8] The first pattern category shown, `[PERSON-NAME]`, is recognizing the name and defining the term as a person's name. The second stretchy pattern category `[GAP]` is merely used to denote a chunk of words that are disregarded.

Table 4. Standard sentences expressed as stretchy patterns.

Original Phrases	Stretchy Patterns
The boy ran quickly to the store today.	The [person] ran [GAP] [time-ref].
The girl ran on the track yesterday.	The [person] ran [GAP] [time-ref].
I tried to sing but my voice cracked terribly.	[first-pron] [GAP] but [GAP] terribly.
He wanted to run but he was terribly sore.	[third-pron] [GAP] but [GAP] terribly.

We engineered a number of SPs. With general and SP features, we trained multiple models, examining them for extracted SPs heavily weighted toward given gangs. Amid thousands of features, several SPs were discernible. To illustrate an SP contribution, consider that models built with general features likely weight heavily toward the Vice Lords those users whose posts contain 'lord' in a reference to the gang and misclassify users whose posts contain 'lord' in reference to a religious entity. A relevant category, [religionVB], with related words, 'praise','worship', and 'exalt', and a corpus-tailored SP, "[religionVB][GAP] the Lord," that characterizes context, serve to tease apart contrasting senses. Modeling excludes new-SP occurrences of 'lord' from unigram calculations and instead matches them to the new SP, which, as a feature, competes for weight as would any new feature. 'Lord' is weighted heavily, as a unigram, in 'proud to be a Lord' and proportionally, as an element of a feature, in 'praise the Lord'. SPs thus enabled us to model error-generating word use variation in the smaller corpus of ~800 users.

5. Analysis

5.1 Qualitative

Due to the large volume of data available from HoodUp, qualitative analysis first explores instances of low and high gang style usage, with a view toward context of occurrence. Alert to changes in frequency, density, variety, and other artifacts of language feature usage, we sought to associate the occurrences, or topic and style *shifts*, with a known function of language, focusing on those related to identity, however defined or determined. We also recognized that usage serving an identity function may co-occur with non-linguistic traits, such as <age>, <location>, and <education>, among others.[9]

5.1.1 Single User Code Switching

The number of variable type categories, variables, instantiations, and combinations thereof soon proved exponentially large and unwieldy. To stabilize the analysis, we found that a *user-driven*

[9] While non-linguistic traits may well be linked to or contribute generally to an individual's identity or specifically to an individual's own sense of their own identity, pursuit of these questions is beyond the scope of the present work.

approach was effective. Tracking an individual user, we neutralized the effects of changes in non-linguistic traits, as individuals usually have single sets of these traits. When we found a forum user whose posts contained a single predominant feature and who then varied style feature instances and combinations as the context changed, we developed hypotheses about the function of identity in gang language and even formulated new research questions. In this report, we focus on the postings of a single user, known in the forum by the username `Valentine GangsterBlood` and show how his use of language shaped our research.

We chose to look at the language of `Valentine GangsterBlood` because he consistently used only one predominant style feature—namely the `S/5` replacement feature—which made it easy to quickly identify his (non-)style usage in posts.[10] He also had an optimal number of total posts, one that was low enough to enable us to read through all of his posts yet high enough for recognition of style usage patterns. As table 5 indicates, out of a total 84 posts, about 69% contained style features, meaning that the majority of his posts contained gang style. As such, examining the contexts of his non-style usage was our first step toward understanding why and how he implemented non-style language. While some non-style posts were merely devoid of opportunities for `S/5` replacement, others were uncharacteristically un-styled and often contained a very definite use of SAE ("`the`" vs. "`da`" for example), which piqued our curiosity.

Table 5. Individual user style.

USER NAME	USER ID	TOTAL POSTS	POSTS WITH STYLE	POSTS W/O STYLE	PERCENTAGE STYLE
`Valentine GangsterBlood`	27288	84	58	26	69%

Although we began by specifically looking at posts where `Valentine GangsterBlood` did *not* use style features, by reading through the entire thread to contextualize the posts, we were able to formalize a description of the codes between which the user switched. The fact that this was even possible reinforced our defining axiom that gang language is a type of code switching. We observed code switching occurring between (1) a range of SAE to AAE vernacular without the `S/5` replacement feature and (2) a range of style usage with the `S/5` replacement feature. We illustrate how the codes contrast with each other in table 6.

Two posts in a single thread by our user are displayed in table 6. The first is longer, SAE-like and contains `S`'s, none of which are replaced with `5`'s. The second is shorter, less standard (note substandard form, "`gonna`") and contains `S`'s, all of which are replaced with `5`'s. While the first post is considered to have some gang style language based on use of gang names, there is no manipulation of the semiotics as there is in the second. Knowing that both posts come from the same thread, it becomes apparent that there is something about `Valentine GangsterBlood`'s language use that is far from arbitrary.

[10] The `S/5` replacement feature is characteristic of gangs affiliated with the *People Nation*, a gang alliance. See table 6.

Table 6. Gang language as code switching.

Username	User ID	Thread ID	Post ID	Post	S/5?
Valentine GangsterBlood	27288	36068	667959	Are there any true ubn members on here or just fake wannabes like soto and Taliban? "when bloods go against bloods they become crabs, because out west only crabs kill crabs, and true homies don't flip on each other."	No
			687632	He*5 gonna* go travel to every*1* ju*5*t to find out who*5* real? Then *5*hot them or *5*tab them if they not__lmao__	Yes

Thread 36068 begins with a user claiming affiliation with the Dirty South Rollin 20 Bloods, a set within the greater Blood gang. While users entering the thread use various Blood-specific style features, Valentine GangsterBlood, who is affiliated with another Blood set, does not use style. His first post is calling out other users for not being real Bloods and asking where the real Bloods are. Several posts later, when he defends his own identity as a Blood gangster, his language contains heavy style usage. There thus appears to be a deliberate use of gang language as a function of identity. Judging from the posts in table 6, when Valentine GangsterBlood feels in control or in power, his need to use the S/5 style feature diminishes; when his identity is attacked, however, the S/5 style feature dominates and functions to assert his identity and belonging.[11]

5.1.2 Use and Non-Use of Gang Language

In virtual spaces, language alone carries the identity function burden. Note that Valentine GangsterBlood's posts *do* usually contain style features, such that, for him, style use is expected or unmarked. With it, he identifies himself as an authority on Blood gang membership. Non-style use for this user, in the context of a thread discussing gang affiliation, would constitute marked or unexpected language use.

Examining forum threads in which Valentine GangsterBlood did not use gang style, we found use and non-use of gang language. This includes threads that we refer to as "informational" as well as instances, in which gang members challenge each other's identities. That is, they call each other out by asking various versions of "who you be" and, in so doing, mimic a similar practice on the streets known as "Where you From?" (Garot, 2007). Here, we briefly touch upon these two phenomena.

What we are calling "informational" are threads in which identity and gang affiliation have low salience and posts exhibit little to no style usage. Valentine GangsterBlood's forum

[11] This finding is important for the Army's use case. Effective exploitation of the CS phenomenon for determining group membership may require identity-related provocation to elicit relevant CS behavior.

activity includes a few of these threads. Two of them pertain to advice on mundane gangster-related topics such as the best light bulb to use to grow marijuana plants indoors and the best way to open a bank account for a business that is a front for some other, most likely illegal, business. Gang identity has low salience in these threads, posts in which rarely incorporate gang style. In a third informational thread, assertion of gang identity remained at a low level as discussion centered on the appearance of advertisements for pornography on the HoodUp Web site. As with the informational threads, gang affiliation carries little weight. Users express annoyance or disgust with the ads, conjecture about reasons for their appearance and suggest ways of blocking them. Perhaps due to the universal nature of these perspectives, their expression takes precedence over that of gang identity and this is reflected in the non-use of gang style in the posts. The other phenomenon we discovered in the forum was the use of various versions of "Who you be" callouts—callouts asking a user to identify their gang affiliation. This reflects the forum users' concern that participants be "real" gangsters. A non-gangster, possibly posing as real, is referred to as a "netbanger." Similar to the instances of "Where You From!" that occur on the street, a phenomenon analyzed by Garot (2007), these callouts are often left ignored. Thus, what appears an appropriate opportunity for forum users to identify their affiliation is flouted in both the online and on the street version of the ritual. Garot (2007) discusses how gang members asked this question on the streets would consider their options and might decide to portray some specific circumstance-dependent identity, a response we have yet to observe online.

5.2 Quantitative

5.2.1 Testing the Predictive Power of Models

Facing enormous volumes of postings and forum users, quantitative analysts train multiple models that predict users' gang affiliations. Recall that, for the affiliation prediction task mentioned in section 4.2.2, analysts extracted features for each user and ran various versions of algorithms on them to determine each user's gang membership. Yet, the non-use of style in certain threads, discussed in section 5.1.2, and by certain users, as discussed in section 4.2.2, suggests that gang style usage can only go so far in determining forum user affiliation. So, analysts compare model accuracies. The first one created was a simple `LogisticRegression+Unigram` model, trained by running the logistic regression algorithm with single feature type, unigram, tokens of which were extracted from the data.[12]

Quantitative analysis crucially includes evaluative examination of the features used to achieve the accuracies. Exploration of feature exploitation by the learning process is possible in the *Explore Results* tab of LightSIDE. With it, analysts can determine how the selected learning algorithms, incorporating the selected options, use the selected features to create the resulting

[12] Names of models consist of the names of the `<LearningAlgorithm>+<OptionSet>` used in training each one. In this model, *stopwords*–or frequently occurring determiners, pro-forms, copulas, and prepositions whose counts may distort the model and skew results–were excluded from the unigrams extracted. There are 118 stopwords in the LightSIDE software, including words such as "the," "is," and "it."

models. Put another way, it clarifies the contributions that <algorithm+options> has each feature make to the model with, for each feature, evaluations of its <frequency>, <average value> and <influence>, which are calculated for the training dataset and displayed.[13]

The Explore Results capability, illustrated in figure 4, displays Classification Results matrix, Feature Table, and Frequency Evaluation for the trained model, LogisticRegression+Unigram. In the upper-central confusion matrix, the cell for actual Crip users who were correctly predicted to be Crips is highlighted and has a black dot adjacent to the count, 388. Note also that the total for the selected cell's row, that is, the number of actual Crips represented in the model training data is 415.[14]

In the upper-right quadrant, note the listing of features and frequencies. Features are associated with users counted in the highlighted central matrix cell and frequencies are feature users from among the same user group. The black-dotted feature in the table is the one being evaluated. Frequencies in neighboring cells are counts of users, from among those represented in the selected matrix cell, in whose language the feature occurred. Thus, 224/415 or ~54 % of actual Crips used the term "cuz."

The model confusion matrix at bottom is an overall Frequency Evaluation of terms selected from the Features Table. Hence the count of feature-incorporating users among correctly classed Crips in the blue cell of that matrix remains 224, as reported in the Table of Feature Frequencies. Sparse distribution of the feature elsewhere in the matrix indicates that the "cuz" feature is heavily weighted toward the Crips. In fact, the data here indicate that, compared with Non-Crips overall, Non-Crips who use "cuz" have a 20% greater likelihood of being misclassified as a Crip.[15]

[13] The LightSIDE software is described generally in sections 2 and 4 of this report.

[14] The model also used the option L1 Regularization, a pre-process for data normalization. Discussion of this option is beyond the scope of this report. Further information about options can be found in the LightSIDE documentation.

[15] The probability (Pr) of Non-Crips (NC) being misclassified as Crips was calculated: msclasNC(99)/totNC(528)=0.19;Pr of NC "cuz" users (NCcuz) being misclassified as Crips: msclasNCcuz(39)/totNCcuz(99)=0.39.

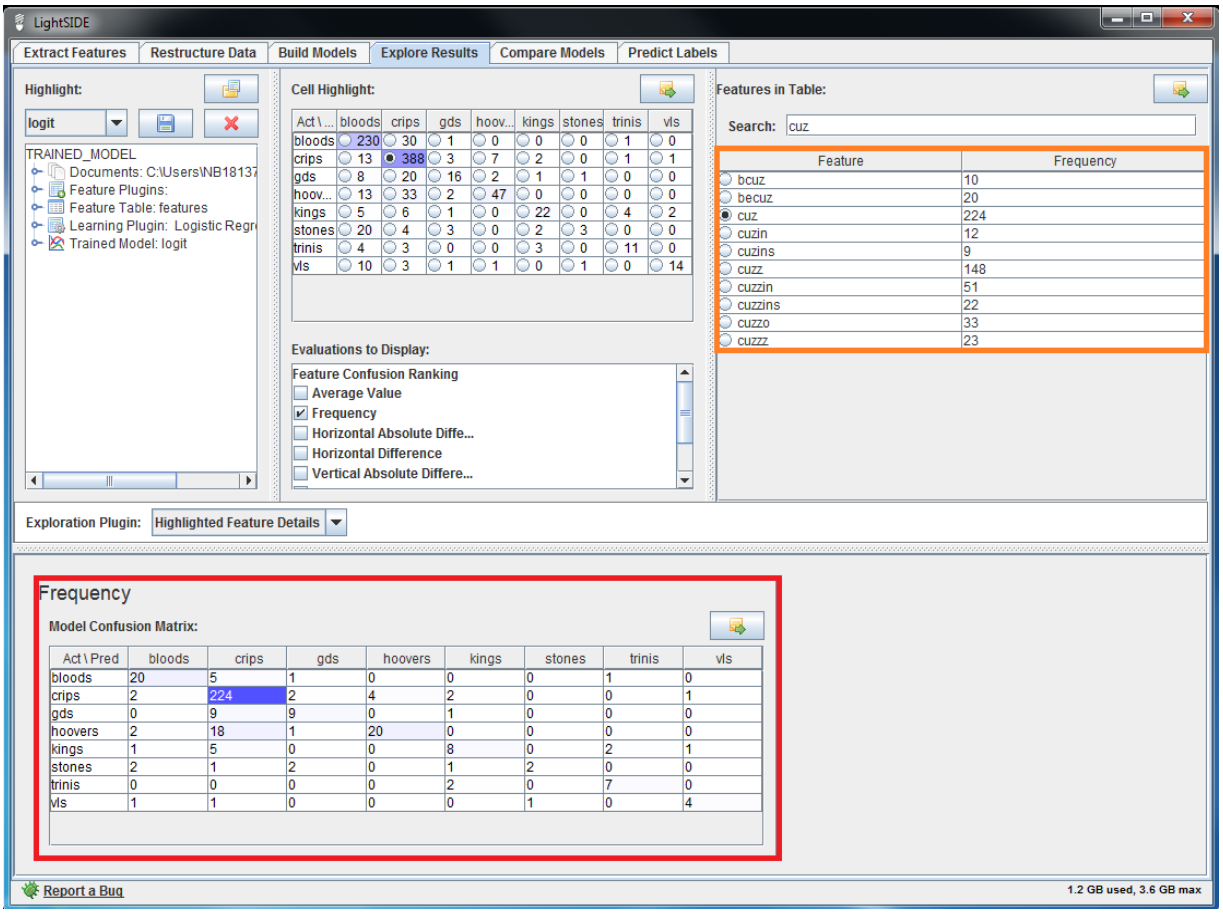


Figure 4. Screenshot of the LightSIDE software's *Explore Results* tab.

Recall that, for the group affiliation question, we were experimenting with a train/test corpus of ~800 non-style-user postings. In section 4.2.2, we noted that an affiliation-relevant aspect of these users' posts was mention of the name of the user's own gang and that sets of terms referring to each gang were assembled and incorporated as features. Thus, while this particular model is based on extraction of unigram features, the `LogisticRegression+Bigram`, `LogisticRegression+BloodFreq` and `LogisticRegression+CripFreq` were based on extracted bigram and two gang name feature sets, respectively. Regardless of model differences, a quantitative analyst will consider suspect any single heavily-weighted feature, such as the unigram "`cuz`." They are indeed, as we have seen here, likely contributors to classification errors. Fortunately, however, the types of errors they create are also good candidates for remediation by features developed from stretchy patterns, described in section 4.2.3.

5.2.2 Correlations with Non-Linguistic Features

Up to this point, our quantitative analytics to support affiliation prediction tasks have taken a *language-internal* text-feature perspective on the *micro level* of the individual user. Yet, there are also *language-external* factors, analyzable at a *macro level*, which contribute to identity,

especially gang identity. Concepts long associated with gangs are trust among individuals and geographic boundary markings. Such links are ingrained in the larger community as well, albeit with more socially accepted instantiations. Terms such as "homies" and "'hoods" have become so established as gangster references to gang members and gang-controlled geographic areas, that SAE speakers may use them familiarly to mean "friends" and "familiar surroundings," respectively. That said, even in an overview such as this one, thoroughness demands consideration of the function that neighborhoods perform for gangs and of variations in neighborhood function as related to a gang's purpose in the area. In this section, we consider regional styles and tendencies associated with gang identity.

Of the many areas in the U.S. that are home to gangs, three key cities, exhibiting sharp contrasts, represent a large percentage of HoodUp users: Chicago, New York, and Los Angeles.[16] Chicago's gang territories, especially on the South Side, tend to be defined by exact streets. Chicago forum users' easy mention of childhood streets that they now defend correlates to great pride in their gang territory and gang identity. In New York, however, gentrification in the Bronx and Brooklyn boroughs has led to increased poverty, and gangs tend less to represent streets where they were raised. Youth are likely to affiliate with gangs in pursuit of elevation rather than from a sense of pride. Los Angeles is comparable to Chicago in that its gang activity predominates in city's south. The gang landscape is quite unlike that of Chicago in the fuzziness of territory boundaries, which—judging from both maps and user posts—creates overlap between gang-controlled regions. Such characteristics contribute to the nature of gang life in each city and can be associated with the demographics of those gang-hosting areas.

Such demographics vary widely. We identified gang culture differences that corresponded with defined regional differences on HoodUp. Online gang maps and posts listing area neighborhoods contributed to our master list of defined neighborhoods for each gang. Having compiled demographic data specific to each neighborhood and its resident gangs, we then consulted the forum to confirm these data with posts.[17]

Chicago's 47% homeownership rate is reflected in its regional gang tendencies. Homeownership is investment in an area. Gang tendencies of well-defined territory boundaries, neighborhood pride, and defense of streets on which members grew up are consistent with and possibly motivated by investment protection. In Brooklyn, the Bronx, and Los Angeles, gang provincial association is weaker and rates of homeownership in the 30%, 20%, and 38% range, respectively, further reinforce this correspondence.

Gang turfs in Chicago can even have neighborhood names as, for example, the Vice Lords' *Lawndale Gardens* area and the Latin Kings' 10-street breakdown known as *Chi Town*. See figure 5. The map snippet in figure 5 is taken from a larger independently labeled map

[16] Compton, technically outside Los Angeles, has a large gang population; its demographics appear later in this section.

[17] Independently created Google Maps defines gang territories in Chicago, Los Angeles, and New York. Virtual mapping's arguable validity motivated forum post corroboration. Online Census Bureau databases provided demographic data by city.

highlighting different gang neighborhoods in South Side Chicago. The full map displays over 115 labeled streets. Note that listed intersections may include names, such as Cal Two One or Coulter LK's, which refer to sets or "gangs within a gang."

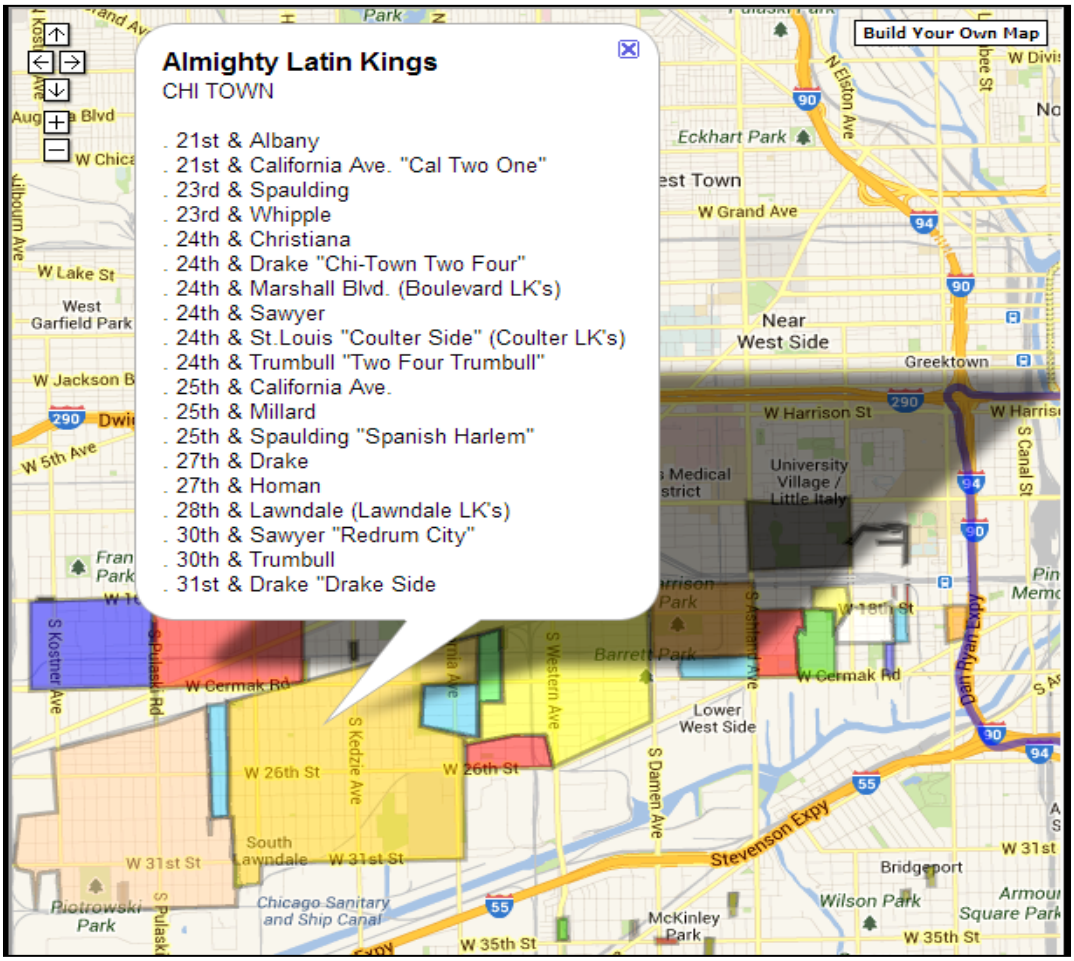


Figure 5. Map of south Side Chicago's Chi Town neighborhood.

As for New York, of the frequently referenced boroughs on HoodUp, i.e., Brooklyn, the Bronx, and Manhattan, demographic associations observed were most pronounced for the Bronx. The 2011 racial breakdown there was White: 11.2%; Black: 43.4%; Asian: 4.2%; and Hispanic/Latino: 53.8%. The latter population ratio is mirrored in the relative percentage of posts referencing Hispanic/Latino gangs, Trinitarios and Latin Kings in the corpus. A key statistic shows 30% of the population living below the poverty line; residents own neither homes nor computers. HoodUp representation by Trinitarios is weak, compared with that of Chicago's ethnically similar gang, the Gangster Disciples.

Southern neighborhoods of Los Angeles are well known for their gang activity, as is the city of Compton, to its southeast. Unlike Chicago's clearly defined and specially named gang territories; those of Los Angeles tend to be poorly demarcated and randomly clustered. Predominantly

African-American in the 60s and 70s, and documented as the birthplace of the gang known as the Bloods, the Compton population's shifting composition in recent years is clear in this 2011 racial distribution—White: 25.9%; Black: 32.9%; Asian: 3.4%; and Hispanic/Latino: 65%. The change in demographics has contributed to increased inter-gang violence, primarily between African-Americans and Hispanic/Latinos, which can affect even the most disinterested bystander and which approaches the level of a racial war.

5. Discussion

Study of gang communication, motivations, and regional demographics is of direct benefit to Army investigations of social meaning in language from both theoretical and methodological perspectives.

This report has suggested that theories of CS and expression of identity can guide analysis of language data to uncover social meanings such as group membership, control, or social power. We found that the use and non-use of a group-specific style by a group member can be expected, and thus indicative of a social meaning. Theories of feature engineering for machine learning, such as the stretchy pattern concept, have also been described. We also set forth a specific classification problem on which they are particularly effective in teasing apart meaningful structural contexts, namely, skewing of learning results due to heavily weighted features.

We exposed methods of exploratory observation of variables, their values, and their relevance to support the formulation of research questions. We showed how user-driven approaches that serve to stabilize variable values and sketch a normative context can be effective in the analysis of and the discovery of social meaning in context-based variation in CS behavior. Methods for experimentation with the automatic exploitation of analytical findings by means of computational modeling and tool interfaces permitting views into an algorithm's manipulation of data features have also been shown.

Both qualitative and quantitative information gleaned about code switching in gangs may well have implications for other "communities within communities," possibly outside the U.S., of Army-strategic importance. Linguistic and computational methodologies established and reported on here may also be adapted for application to Army-relevant situations, datasets or their development, and modeling tasks.

6. Next Steps in Ongoing Work

As a *contribution* to ongoing work, the methodology described here is a first pass at a two-pronged, qualitative-quantitative approach to addressing the discovery and exploitation of social meaning, such as group membership and identity, in CS behavior in the online HoodUp gang forum. As a work in progress, decisive conclusions about connections between gang language and gang members' identity are, by definition, yet to be formulated. We have enough evidence to show a positive association between language use on the forum and identity. While still too early for overarching statements of findings, what we do have are several plausible explanations that apply to individual users and that can serve as preliminary hypotheses for testing.

We will continue analyzing user posts and user variables for consistent patterns in language and identity across the HoodUp forum. Specifically, we will find diverse categories of "informational" threads to determine whether identity consistently maintains a low salience and if correlatively style features remain low in these posts. In ongoing analysis of "Who you be" callouts, we will attend to callout responses for comparison with street responses, as online unresponsiveness seems to serve the same purpose as the street-version quip, "I don't bang." Posts in a variety of threads will serve to confirm or negate observational hypotheses made on the basis of individual users' language, with possible hypothesis adjustment, to detail linguistic underpinnings of gang forum posts that relate to identity, according to Myers-Scotten's markedness model. Confirmed hypotheses will then be applied to code-switched African language social media data in Zulu and Swahili to support the Army's needs and to understand how identity in code switching can serve positively to impact relevant Army scenarios.

Respected analytic tools such as the LightSIDE software will continue to play a significant role in ongoing project work. We plan to contribute to development of such enabling analytic tools and to explore relevant capabilities, such as annotation, of existing freeware, such as the General Architecture for Text Engineering (GATE).[18] Exploiting and building on the tool functionalities, we improve and expand on the social-meaning-producing linguistic analysis, both computationally and in the formulation of relevant research questions. The numerous techniques that contributed to analysis of the gang data, to include the preprocessing strategies, the data displays for discovery and data structures for engineering of effective features, the annotation categories, algorithms, options, and complex indexing of the Chive infrastructure are adaptable to new datasets in different languages and to new classifications of language features. Our gang demographics research has reached a stage where, among possible next steps in classification, could easily figure geographic categories, such as "neighborhood," which would build upon our

[18] General Architecture for Text Engineering Web site. http://gate.ac.uk.

implicit tracking of language use and demographic associations between physical and virtual cultures.

Tuning techniques for Army-relevant situations, language, and locations will be particularly valuable in support of the Army's social meaning in code switching in African languages. Analyzing the corpus of Zulu-English and Swahili-English code-switched postings collected from Facebook and YouTube, we plan to use the GATE software for identifying letter combinations that are unique to Swahili. With these annotated, we can use LightSIDE to train models for identifying code switches in new Swahili posts and use the fact, frequency, or type of code switching as features in classifying users for trustworthiness. We also plan to incorporate SP structuring for the annotated Swahili data. With GATE, we are creating the categories, `<discourse marker>`, `<stopword>`, and `<general>` Swahili term, and annotating instances of these categories in the corpus. These manually created categories, LightSIDE pre-prepared word categories and general features, such as `<unigram>`, can then be used in model building, error analysis, and SP feature discovery and engineering.

As there is ample data for analysis, we are building on the gang demographics research. The team plans to build Latent Dirichlet Allocation (LDA) models (Blei, Ng, and Jordan, 2003) based on neighborhoods, so as facilitate analysis of HoodUp users' regional linguistic tendencies. Resources enabling the LDA experiments include a program for randomizing the training and testing datasets, with attention paid to ensuring equivalent proportional representation of each gang in the two datasets. These approaches and the enabling tools are also applicable to many types of data. Next steps in our ongoing work include learning to apply them to data that contributes to Army solutions.

7. References

Adams, K. L.; Winter, A. Gang Graffiti as Discourse Genre. *Journal of Sociolinguistics* **1997**, *1* (3), 337–360.

Blei, D.; Ng, A.; Jordan, M. Latent Dirichelet Allocation. *Journal of Machine Learning Research* **2003**, *3*: 993–1022.

Conquergood, D. Homeboys and Hoods: Gang Communication and Cultural Space. In C.R. Huff (Ed.) *Group Communication in Context: Studies of Natural Groups* (pp. 23–55). Hillsdale, NJ: Lawrence Erlbaum Associates. 1994.

Gianfortoni, P.; Adamson, D.; Rosé, C. Modeling Stylistic Variation in Social Media with Stretchy Patterns. *Proceedings of First Workshop on Algorithms and Resources for Modeling Dialects and Language Varieties* 2011.

Haugen, E. The Analysis of Linguistic Borrowings. *Language* **1953**, *26*,210–231.

HoodUp Web site. http://www.thehoodup.org. Not accessed.

Jeblee, S.; Piergallini, M.; Rodriguez, A.; Vaughn, C. *Gang Language in Online Forums*. Unpublished manuscript, Carnegie Mellon University, Pittsburgh, PA, 2013.

Mayfield, E.; Rosé, C. P. LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts. In Shermis, M. and J. Burstein, Eds. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, New York, NY: Routledge, Academic Press, 124, 2013.

Mayfield, E.; Rosé, C. P. LightSIDE Text Mining and Machine Learning User's Manual. http://www.lightsidelabs.com. (retrieved July 3, 2013).

Myers-Scotton, C. The negotiation of identities in conversation: A theory of markedness and code choice. *International Journal Sociology of Language* **1983**, *44*,116–136.

Nguyen, D.; Rosé, C.P. Language Use as a Reflection of Socialization in Online Communities. *Proceedings of the Workshop on Language in Social Media* (pp. 76–85), 2011.

List of Symbols, Abbreviations, and Acronyms

AAE	African American English
ARO	U.S. Army Research Office
CMU	Carnegie Mellon University
CS	code switching
CSV	comma separated value
GATE	General Architecture for Text Engineering
HU	Howard University
LDA	Latent Dirichlet Allocation
LTI	Language Technologies Institute
PIRT	Partnership in Research Transition
SAE	Standard American English
SPs	stretchy patterns
SQL	Structured Query Language
SVM	support vector machine

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) GOVT PRNTG OFC A MALHOTRA

2 (PDFs) DIRECTOR US ARMY RESEARCH LAB ATTN RDRL-CIO-LL IMAL HRA MAIL & RECORDS MGMT

2 (PDFs) US ARMY CERDEC CP&I ATTN RDER-CPM-IM R SCHULTZ ATTN RDER-CPM-IM T TRUONG

2 (PDFs) US ARMY RESEARCH LAB ARO RESEARCH TRIANGLE PARK ATTN RDRL-ROI-O PATRICIA HUFF ATTN RDRL-ROI-M JOHN LAVERY

11 (PDFS) US ARMY RESEARCH LAB
ATTN RDRL-CII-T
M VANNI
S LAROCCA
C VOSS
S TRATZ
M HOLLAND
ATTN RDRL-CII-B
R WINKLER
L TOKARCIK
ATTN RDRL-SES-A
L KAPLAN
ATTN RDRL-CII
B BROOME
R HOBBS
ATTN RDRL-DP
V EMERY

4 (PDFS) US ARMY RESEARCH LAB
ATTN RDRL-CII-C-M
M THOMAS
E BOWMAN
S KASE
H ROY

2 (PDFs) CMU LTI, TELEDIA LAB GATES-HILLMAN CENTER 5415 ATTN DR. CAROLYN ROSE ATTN DR. LORI LEVIN

2 (PDFs) HU, DEPT OF ENGLISH LOCKE HALL 232 ATTN DR. NKONKO KAMWANGAMALU ATTN MS BARBRA E CHIN, student

2 (PDFs) HU CEACS, DEPT OF ELEC & COM ENGG L.K. DOWNING HALL 1922 ATTN DR. MOHAMED CHOUIKA ATTN MS CANDACE A ROSS, student

INTENTIONALLY LEFT BLANK.